This is a previous version of the article published in Social Indicators Research. 2018, 140(3): 1279-1300. doi:10.1007/s11205-017-1818-7

The balanced worth: A procedure to evaluate performance in terms of ordered attributes

Carmen Herrero
University of Alicante & Ivie
&
Antonio Villar
Universidad Pablo de Olavide & Ivie

Abstract

There are many problems in the social sciences that refer to the evaluation of the relative performance of some populations when their members' achievements are described by a distribution of outcomes over a set of ordered categories. A new method for the evaluation of this type of problems is presented here. That method, called balanced worth, offers a cardinal, complete and transitive evaluation that is based on the likelihood of getting better results. The evaluation of a population is based on the probability of obtaining better results for an agent of this population than for an agent of another. The balanced worth is a refinement of "the worth" (Herrero & Villar (2013)) that overcomes its excessive sensitivity to the differences, due to the presence of ties. We also discuss how this method can be applied for the case of heterogeneous populations and show how it can be applied in different contexts. An empirical example, regarding life satisfaction in Spain is used to illustrate the working of this method.

Key words: Evaluation method, categorical variables, relative group performance.

Address for correspondence: Antonio Villar, Department of Economics, Universidad Pablo de Olavide, Ctra. Utrera km. 1, 41013 Seville, Spain.

Email address: avillar@upo.es

Carmen Herrero: cherreroblanco@gmail.com

Acknowledgements: Thanks are also due to Héctor García Peris, for his help in developing the algorithm that computes the evaluation, and to an anonymous referee for very helpful comments and suggestions.

1 Introduction

The purpose of this paper is to present a methodology to evaluate the relative performance of social groups when their achievements are described by a distribution of outcomes over an ordered set of categories.

Let us present three evaluation problems that will help understanding the type of situations where this methodology applies.

Evaluation problem 1: Clients' satisfaction

A hotel chain is willing to assess the degree of satisfaction of its clients in the different hotels it runs in a given country. Customers are asked to fill a simple questionnaire in which they report on their satisfaction with the services provided by the hotel, using a five level scale that runs from "highly satisfied" to "highly unsatisfied". The general manager wants to know how the different hotels of the chain perform in order to device an incentive scheme. The informational inputs are the distributions of the clients of each hotel on the five possible levels of satisfaction.

Evaluation problem 2: Educational achievements

The *Programme for International Students Assessment* (PISA) is a project coordinated by the OECD that evaluates the educational achievements of fifteen-year old students in more than sixty countries, regarding reading comprehension, mathematics and science. The basic results correspond to the scores of a test that each student in the sample performs. The OECDE defines six *levels of proficiency* that summarize the tasks that students are expected to manage. The distribution of the population of the different countries along those levels of competence is rather different, even when the average values of the test scores are similar. It is therefore interesting to

compare the countries educational achievements in terms of the distribution of competences among those six levels.

Evaluation problem 3: Intellectual influence

An international research association is willing to compare the academic performance of the countries that are represented in the association, in terms of their scientific publications. The key informational input for such an evaluation is the distribution of the scientific publications into a set of categories that define the levels of relevance of those publications. The standard way of defining those categories is by taking a partition of the world distribution of the citations in the discipline (e.g. deciles) and comparing the distributions of the countries' publications into those categories.

These examples illustrate the diversity of the evaluation problems we consider and its relevance in social sciences. Needless to say, the evaluation will be much more useful if we are able to produce not only a ranking of those populations but also a quantitative measure of their performance. This is our goal: providing a cardinal measure of relative performance for distributions of ordered qualitative variables.

The most usual ways of evaluating this type of problems is either by recurring to some notion of *stochastic dominance* or to some *scoring rule* that attaches weights to the different categories and evaluates performance in terms of weighted averages. Both evaluation methods present some inconveniences. On the one hand, stochastic dominance typically yields a partial ordering, so that we can only rank some distributions, but not all. Furthermore, there is no cardinal evaluation associated to that partial ranking (that is, we can only say if one distribution is better than other, but not how much better).–On the other hand, scoring rules can be very arbitrary and the

evaluation results may be too much dependent on the scores attributed to the different categories.

We present here a new evaluation method for this type of problems, called the ***balanced worth***, that is cardinal, complete and transitive, and does not involve any external weighting scheme. The evaluation of each population is based on the likelihood that a representative member gets better results than a representative member of another population. This way of evaluating distributions of qualitative variables appears in Lieberson (1976), for the case of two-population problems. The concept of *worth* was introduced in Herrero and Villar (2013) as a transitive extension of this notion to an arbitrary number of populations. Yet this extension does not compute the probability of ties when comparing the results of the groups' representative agents. That implies that the associated evaluations take only into account the parts in which the distributions differ and ignore the parts in which they are similar (see discussion below).

This paper introduces the balanced worth as a refinement of the worth that overcomes its excessive sensitivity to the differences in the outcome distributions, by taking into account the probability of ties. It also provides an intuitive explanation of this evaluation method, analyses its nature and properties, includes some extensions, discusses how it relates to the worth, and shows how can be applied in different contexts.

Related evaluation criteria appear in a variety of problems, such as the statistical measure of distributional similarities (Li, Yi and Jestes 2009, Martínez-Mekler *et al* 2009, Gonzalez-Diaz, Hendrichx and Lohmann 2013), the ranking of income distributions in different contexts (Shorrocks 1983, Bellú and Liberati 2005, Bourguignon, Ferreira and Leite 2007, Yalonetzky 2012, Sheriff and Maguire 2013, Cuhadaroglu 2013), the analysis of segregation and discrimination (Reardon and

Firebaugh 2002, Grannis 2002, Echenique and Fryer 2005, Chakravarty and Silber 2007, Frankel and Volij 2011), the evaluation of scientific influence (Pinski and Narin 1976, Laband and Piette 1994, Palacios-Huerta and Volij 2004, Crespo, Li and Ruiz-Castillo 2013), the comparison of network structures (Rosvall and Bergstrom 2007), or the allocation of scores in tournaments (Laslier 1997, Slutzki and Volij 2006).

The paper is organised as follows. Section 2 describes the evaluation method and its main properties, including a theorem where the existence and uniqueness of the evaluation is ensured. It also provides an empirical illustration regarding life satisfaction in Spain by age groups. Section 3 deals with the applicability of this evaluation method to the case of heterogeneous populations (each group consists of different *types*). It takes up the former empirical illustration to extend the analysis when the population of each age group consists of two different types, men and women. Section 4 is devoted to the discussion of the balanced worth vis a vis the worth. We re-evaluate the empirical analysis in the original paper by Herrero and Villar (2013) in order to show their differences. Section 5 closes the paper with a discussion of the main features of this evaluation method and a short description of some fields of application.

2 The model

2.1 The reference problem and the evaluation method

The reference problem consists of evaluating the relative performance of a collection of g populations, $G = \{1, 2, \ldots, g\}$, whose achievements are described by a distribution of values over a finite set of categories that are linearly ordered (ordinal categorical variables). Those populations, also called ***groups***, are to be understood as related in some way, e.g. they correspond to subsets of a larger set, such as the plants of

a firm, the regions of a country or the countries of an association. This is so because otherwise making a relative comparison makes little sense.

Each population $i \in G$ has n_i elements, also called ***members***. Associated with each element there is a value that measures individual performance, referred to as ***outcomes***, which we assume can take on a finite number of values,[1] called ***levels***. We assume that those levels are ordered from best to worse. That is, level 1 is better than level 2, level 2 is better than level 3, etc.

A ***distribution*** of outcomes for population i is a vector $\mathbf{a}(i) = (a_{i1}, a_{i2}, ..., a_{is})$ that describes the fraction of its members into each admissible level of performance. That is, $a_{ir} = n_{ir} / n_i$, where n_{ir} is the number of elements in population i with outcome level r. Clearly, $a_{ir} \geq 0, \sum_{r=1}^{s} a_{ir} = 1$.

An ***evaluation problem***, or simply a ***problem***, refers to the comparison of the relative performance of those populations in terms of the behaviour of their members. That is, assessing the relative goodness of the distributions $\mathbf{a}(1), \mathbf{a}(2), ..., \mathbf{a}(g)$. An evaluation problem can thus be summarized by the matrix **A** made of all those **a**(i) distributions, which we interpret as the rows of **A**.

The basic principle to compare the populations' performance refers to the probability of getting better outcomes. For a given problem **A** we denote by p_{ij} the probability that a member chosen at random from population i exhibits a higher level of performance than a member chosen at random from population j. As the levels are ordered from best to worst, we can calculate that probability as follows:

$$p_{ij} = a_{i1}(a_{j2} + ... + a_{js}) + a_{i2}(a_{j3} + ... + a_{js}) + ... + a_{i(s-1)}a_{js}$$

Let $e_{ij} = e_{ji}$ stand for the probability that a member of group i exhibits the same level of performance than a member of group j. By construction, we have:

$1 = p_{ij} + p_{ji} + e_{ij}$.

We now describe a procedure to get quantitative estimates of the relative desirability of those distributions of ordered categorical data. This procedure can be described in terms of *a contest* in which each group is confronted randomly with some other.

The simplest case: two groups

Suppose we have just two groups, *i* and *j*. In order to determine which group exhibits a better distribution of outcomes, we propose the following protocol. One member from each group will be selected at random and they will be confronted. If the member from group *i* beats that from group *j* (that is, it exhibits a higher level of performance), then the distribution of group *i* is declared *better than* that of group *j*, and vice-versa. In case both members exhibit the same level of performance, a coin is tossed and each group will be declared best with probability ½.

The probability that group *i* be declared better than group *j* is given by:

$$p_{ij} + \left(e_{ij} / 2\right)$$

That is, the probability that *i* beats *j* plus half the probability of a tie. Similarly, the probability that group *j* be the winner in this confrontation is:

$$p_{ji} + \left(e_{ji} / 2\right)$$

Given these data, how should we value the outcomes of those two groups? Our proposal is simple and natural: *make the value of each group proportional to the probability of being a winner*. That is, if we call w_i, w_j the evaluations of those two groups, we have:

$$\frac{w_i}{w_j} = \frac{p_{ij} + \left(e_{ij} / 2\right)}{p_{ji} + \left(e_{ij} / 2\right)} \qquad [1]$$

We call ***proportionality*** to this evaluation principle. Note that this formula has one degree of freedom, so that we can choose units arbitrarily. For the case of two groups, therefore, the proportionality principle fully determines the evaluation formula, except for the choice o units.

Equation [1] can be rewritten as:

$$w_i = \frac{\left[p_{ij} + \left(e_{ij} / 2\right)\right] w_j}{p_{ji} + \left(e_{ji} / 2\right)} \qquad [1']$$

In this way the evaluation of group *i* appears as the ratio of two interesting expressions. The one in the numerator can be regarded as the ***relative advantage*** of *i* over *j*, as it corresponds to the probability of getting better outcomes, weighted by the evaluation of group *j*. The denominator can be seen as the ***relative disadvantage*** of group *i* with respect to population *j*, as it expresses the probability of getting worse outcomes.

The general case: g ³ 2 groups

It is easy to check that if we apply this criterion for pair-wise comparisons when there are more than two groups, we may find a cycle because the evaluation they induce is not transitive. The example in Table 1 illustrates this problem. It describes the outcome distribution of three groups, 1, 2 and 3, into four levels of performance, I, II, III and IV.

Table 1 around here

We find that: (i) $p_{12} + (e_{12} / 2) = 0.525$, $p_{21} + (e_{21} / 2) = 0.475$, which implies that group 1 is better than group 2. (ii) $p_{23} + (e_{23} / 2) = 0.525$, $p_{32} + (e_{32} / 2) = 0.475$,

which implies that group 2 is better than group 3. And (iii) $p_{31} + (e_{31}/2) = 0.585$, $p_{13} + (e_{13}/2) = 0.415$, which implies that group 3 is better than group 1, thus creating a cycle.

The simplest way of avoiding this problem is by taking expectations. This may be regarded as applying the same process as in the case of two groups, but now choosing also randomly the group with which group *i* will be compared. That is, the value of group *i* will be given by the following formula:

$$w_i = \frac{\frac{1}{g-1}\sum_{j \neq i}\left(p_{ij} + (e_{ij}/2)\right) w_j}{\frac{1}{g-1}\sum_{j \neq i}\left(p_{ji} + (e_{ji}/2)\right)}, \qquad i, j = 1, 2, \ldots, g \qquad [2]$$

This expression is a generalization of equation [1']. Now the numerator describes the ***average relative advantage*** of the distribution of population *i* with respect to the rest, whereas the denominator corresponds to the ***average relative disadvantage*** of population *i* with respect to the rest. Trivially, equation [2] collapses to equation [1'] when there are only two populations.

The balanced worth

Equation [2] thus provides a complete and transitive extension of the proportionality principle in equation [1]. We call **balanced worth** to this evaluation method, because it is a refinement of the concept of worth introduced in Herrero and Villar (2013) (see the discussion below). Note that calculating the balanced worth in the general case requires solving the following simultaneous equation system:

$$\left.\begin{array}{l} w_1 \sum_{j\neq 1}\left[p_{j1}+\left(e_{j1}/2\right)\right]=\sum_{j\neq 1}\left[p_{1j}+\left(e_{1j}/2\right)\right]w_j \\ w_2 \sum_{j\neq 2}\left[p_{j2}+\left(e_{j2}/2\right)\right]=\sum_{j\neq 2}\left[p_{2j}+\left(e_{2j}/2\right)\right]w_j \\ \ldots \quad \ldots \quad \ldots \quad \ldots \\ w_g \sum_{j\neq g}\left[p_{jg}+\left(e_{jg}/2\right)\right]=\sum_{j\neq g}\left[p_{gj}+\left(e_{1j}/2\right)\right]w_j \end{array}\right\} \qquad [3]$$

We discuss in the following sections the nature and implications of this evaluation formula. Before that let us formally state that a solution to the system of g equations with g unknowns [3], always exists.

Theorem 1: *Let* **A** *be an evaluation problem. Then:*

(i) There exists a vector $\mathbf{v}^* \in \Re_+^g$ *that solves equation system [3]. That is, a vector* **v*** *such that:*

$$v_i^* = \frac{\sum_{j\neq i}\left(p_{ij}+\frac{e_{ij}}{2}\right)v_j^*}{\sum_{j\neq i}\left(p_{ji}+\frac{e_{ij}}{2}\right)}, \quad i=1, 2, ..., g$$

(ii) If $\left(p_{ij}+\frac{e_{ij}}{2}\right)>0 \ \forall \ i, j$ *then the solution is unique (up to a scalar multiplication) and strictly positive.*

Proof

(i) Let $V=\left\{\mathbf{v}\in \Re_+^g \ / \sum_{i=1}^{g} v_i = g\right\}$ and consider the function $\varphi : V \to \Re$, given by:

$$\varphi_i(\mathbf{v}) = v_i - \frac{1}{g-1}\left(v_i \sum_{j\neq i}\left(p_{ji}+\frac{e_{ij}}{2}\right) - \sum_{j\neq i}\left(p_{ij}+\frac{e_{ij}}{2}\right)v_j\right)$$

As $\sum_{j\neq i}\left(p_{ji}+\frac{e_{ij}}{2}\right)\leq g-1$, we have:

$$\varphi_i(\mathbf{v}) \geq v_i - v_i + \frac{1}{g-1}\sum_{j\neq i}\left(p_{ij}+\frac{e_{ij}}{2}\right)v_j \geq 0$$

Moreover,

$$\sum_{i=1}^{g} j_i(\mathbf{v}) = g - \frac{1}{g-1}\left(\sum_{i=1}^{g} v_i \sum_{j\neq i}\left(\left(p_{ji} + \frac{e_{ij}}{2}\right)\right) - \sum_{i=1}^{g}\sum_{j\neq i}\left(p_{ij} + \frac{e_{ij}}{2}\right)v_j\right)$$

Note that, by construction,

$$\sum_{i=1}^{g} v_i \sum_{j\neq i}\left(\left(p_{ji} + \frac{e_{ij}}{2}\right)\right) = \sum_{i=1}^{g}\sum_{j\neq i}\left(p_{ij} + \frac{e_{ij}}{2}\right)v_j$$

which means that $\sum_{i=1}^{g} j_i(\mathbf{v}) = g$ and hence that function j maps V into itself. As it is a continuous function and V is a compact convex set, Brouwer's Theorem (e.g. Zeidler (1986)), ensures the existence of a fixpoint, $\mathbf{v}^* = j(\mathbf{v}^*)$. That is,

$$v_i^* = v_i^* - \frac{1}{g-1}\left(v_i^* \sum_{j\neq i}\left(p_{ji} + \frac{e_{ij}}{2}\right) - \sum_{j\neq i}\left(p_{ij} + \frac{e_{ij}}{2}\right)v_j^*\right)$$

and, therefore,

$$v_i^* = \frac{\sum_{j\neq i}\left(p_{ij} + \frac{e_{ij}}{2}\right)v_j^*}{\sum_{j\neq i}\left(p_{ji} + \frac{e_{ij}}{2}\right)}, \quad i = 1, 2, ..., g$$

(ii) Assume now that $\left(p_{ij} + \frac{e_{ij}}{2}\right) > 0$ " i, j. Then, the solutions must be strictly positive.

To prove uniqueness, suppose there are two strictly positive vectors, **w**, **y**, that solve the equation system [3]. Then, we can write:

$$\sum_{j\neq i}\left(p_{ji} + \frac{e_{ij}}{2}\right) = \frac{\sum_{j\neq i}\left(p_{ij} + \frac{e_{ij}}{2}\right)w_j}{w_i} = \frac{\sum_{j\neq i}\left(p_{ij} + \frac{e_{ij}}{2}\right)y_j}{y_i}, \quad i = 1, 2, ..., g$$

For a given i, this expression can be rewritten as:

$$A = \sum_{i=1}^{g-1} B_i x_i = \sum_{i=1}^{g-1} B_i z_i$$

where all terms are strictly positive, with $x_j = w_j / w_i$, $z_j = y_j / y_i$. But this is the equation of a hyperplane with a given normal, which means that vectors **x** and **z** are to be proportional. That is, the solution is unique up to the choice of units.

Q.e.d.

We call **balanced worth vector** to that solution $\mathbf{v}^* = \left(v_1^*, ..., v_g^*\right)$. The balanced worth attaches to each group the ratio between the average relative advantage of that group and the average relative disadvantage. It is, therefore, a rather intuitive evaluation procedure.

The balanced worth satisfies the standard requirements of any evaluation function. In particular:

- *Anonymity*: the evaluation only depends on the individuals' performance and not on other aspects such as labels or names. Therefore, permuting the realizations between the agents will not change the evaluation.
- *Symmetry*: if two groups have identical distributions, $\mathbf{a}(i) = \mathbf{a}(j)$, then their corresponding balanced worth values will be the same.
- *Monotonicity*: if the members of group *j* improve their outcomes whereas all other groups outcomes remain unaltered (that is, the distribution of group *j* shifts to the upper levels of performance), then the balanced worth of group *j* will increase. This, in turn, implies *stochastic dominance*: If the distribution of one group stochastically (first order) dominates the distribution of another, then it will exhibit a larger balanced worth.

The computation of the balanced worth can be directly obtained through a friendly and freely available algorithm, hosted in website of the *Instituto Valenciano de Investigaciones Económicas* (Ivie). The required address is: http://www.ivie.es/balanced-worth/ . This webpage explain how this algorithm works (it computes the dominant eigenvector of a suitable Perron matrix) and how to proceed to implement the calculations. In particular, the balanced worth can be obtained directly

from the matrix of relative frequencies that can be plugged into the algorithm as an excel table, thus saving much time and effort. By default the algorithm normalizes the values of the balanced worth making the mean of the groups equal to 1.

2.2 An empirical illustration: Life satisfaction in Spain

Let us illustrate the working of this evaluation method by considering the problem of assessing life satisfaction in Spain.

During 2013 the European Union (EU) elaborated for the first time a comparative study regarding the member states' quality of life, from a subjective perspective (see Eurostat 2015). The data were collected through the 2013 Ad-hoc module of EU SILC on subjective well-being. Life satisfaction is one of the three dimensions that define subjective well-being, based on an overall cognitive assessment (the other two being *affects* and *eudaimonics*). Life satisfaction represents how a respondent evaluates life as a whole, that is, an assessment comprising all areas of a person's existence. It focuses on how people are feeling "these days" rather than specifying a longer or shorter time period (see Veenhoven (1991, 3), Pavot and Diener (2008, 137)). Economists may think of that as a measure of individual welfare.

Life satisfaction is measured on a 0-10 scale (where 0 is "not satisfied at all" and 10 "fully satisfied"). To facilitate analyses, those numerical evaluations were grouped into different categories, according to the statistical distribution of the answers. In the case of Spain, the National Statistical Office (INE) used four categories that we term: Low (0-4 points), Fair (5-6 points), High (7-8 points) and Very high (9-10 points). Table 2 provides the distribution of answers by different age groups, together with the balanced worth (normalised so that the mean equals 1) and the normalised means of the different age groups (global mean equals 1).

Table 2 about here

The most obvious message of these data is that life satisfaction diminishes with age. More interesting is the comparison between the balanced worth, which computes the differences in the distributions by age groups, and the (normalized) means. Even though both measures exhibit a decreasing pattern with age, the differences by age groups are much larger in the case of the balanced worth. Indeed, the coefficient of variation of the balanced worth values is almost four times that of the mean values (0.153 versus 0.04).

4. Heterogeneous populations

An implicit assumption of the evaluation model just described is that groups are homogeneous, so that the distribution of the outcome variable is the sole relevant information. Yet, when groups are heterogeneous, one might be interested in evaluating not only the observed outcomes but also the extent to which those outcomes reflect diverse structural characteristics of the groups that affect the agents' performance. Aspects such as sex, race, age, nationality, parental background, or wealth, can influence individual outcomes in particular problems and it is interesting to know to what extent the observed outcome differences correspond to differences in the composition of the groups.

There is a number of related but different questions that can be addressed when dealing with heterogeneous populations and the use of the balanced worth has to be adapted to each case. Think, for instance, we are evaluating perceived health in the OECD countries. Each individual in the sample rates her perceived health in one out of five different health states, ranging from very good to very bad. If we identify each OECD country with a group and apply the balanced worth to evaluate the health state of

those countries, we are disregarding the fact that part of the observed differences in the distribution of responses reflects differences in the demographic composition of the populations. And there is strong evidence that health perceptions are age dependent.

How to address this problem mostly depends on the type of comparisons deemed relevant. One possibility is considering each population subgroup as a different group, so that the evaluation is made with respect to the $t \times g$ subgroups, where t is the number of different types within each group. We call this the ***joint evaluation***. In the example of the health states, this means that we think relevant comparing the health status of young people in France relative to old people in Germany, say. Another possibility is that of making comparisons among population subgroups with similar characteristics (e.g. the health states of the young in all countries). We shall refer to those comparisons as the ***separate evaluation by types***. It provides an evaluation of the *between groups* relative performance by types. Still a different evaluation problem in the context of heterogeneous populations refers to the evaluation of the degree of heterogeneity within the groups. In the health example that amounts to evaluate how different are the results on perceived health between generations in different countries. We call this ***separate evaluation by groups***. This evaluation provides a measure of *within group* heterogeneity.

Which form of comparison is more adequate depends on the problem at hand and it is part of the modelling choices open to the researcher. We shall now describe briefly how to deal with those questions.

3.1 Separate evaluation by types

The evaluation problem in the case of heterogeneous populations can be framed as follows. We have, as before, an evaluation problem involving g groups whose

achievements regarding some aspect are given in terms of *s* ordered levels. The novelty now is that the population of each of those g groups can be classified in terms of t different ***types***, indexed by $t = 1, 2, \ldots, t$. Each type within a group gathers those members with similar characteristics, so that different types correspond to differential structural traits in the population of that group. In the example regarding perceived health the types are usually defined by age intervals (e.g. young, adult and old), so that the implicit assumption is that all agents in the same age interval are directly comparable in terms of their health states.

The outcome of each group $i = 1, 2, ..., g$ will now be described by a collection of t distributions, $\mathbf{a}^t(i) = \left(a_{i1}^t, a_{i2}^t, ..., a_{is}^t\right)$, for $t = 1, 2, \ldots, t$ (a contingency table). Each term $a_{ir}^t = \frac{n_{ir}^t}{n_i^t}$ corresponds to the share of the population of type *t* within group *i* with level of achievement *r*. Here n_{ir}^t, n_i^t are the number of members of group type *t* with level *r* within group *i*, and the total number of members of type *t* within group *i*, respectively. For all $t = 1, 2, ..., t$, all $i = 1, 2, ..., g$, we have: $\sum_{r=1}^{s} a_{ir}^t = 1$.

We can now evaluate the relative performance of each type among the groups (e.g. comparing health states between old people across countries), by considering the evaluation problem defined by the following collection of $(g \times s)$-matrices:

$$\mathbf{A}(t) = \begin{bmatrix} \mathbf{a}^t(1) \\ \mathbf{a}^t(2) \\ ... \\ \mathbf{a}^t(g) \end{bmatrix}, \quad t = 1, 2, \ldots, t \qquad [3]$$

The balanced worth of each of those problems, $\mathbf{w}(t)$, $t = 1, 2, ..., t$, tells us about the relative performance of the corresponding type across groups. The implicit assumption is that comparing the outcomes of different types is not relevant.

The overall evaluation of the group can be obtained as a weighted average of those types, with weights corresponding to the population shares That is,

$$W_i(\mathbf{A},t) = \sum_{t=1}^{t} \frac{n_i^t}{n_i} w_i(t) \qquad [4]$$

Each term of this sum in equation [4] is the product of two numbers. The first one is the share of type *t* in the group and reflects its ***composition***. The second evaluates the performance of type *t* in this group relative to type *t* members of other groups. It provides a measure of the ***return*** of the type in this group, relative to the return of the same type in other groups.

We can now estimate the composition effect by comparing that value in equation [4] with one in which the composition of group *i* corresponds to a given standard, $W^C(.)$. Suppose that we take the average composition of the groups as the standard, for the sake of simplicity. That yields,

$$W_i^C(\mathbf{A},t) = \sum_{t=1}^{t} \frac{\sum_{i=1}^{g} n_i^t}{\sum_{i=1}^{g} n_i} w_i(t)$$

The composition effect will thus be measured by:

$$C(\mathbf{A},t) = W_i(\mathbf{A},t) - W_i^C(\mathbf{A},t) = \sum_{t=1}^{t} \left(\frac{n_i^t}{n_i} - \frac{\sum_{i=1}^{g} n_i^t}{\sum_{i=1}^{g} n_i} \right) w_i(t) \qquad [5]$$

Similarly, we may be willing to calculate the effect of the differential returns of the types by comparing [4] with some standard. If we choose the average return, we would have:

$$W_i^R(\mathbf{A},t) = \sum_{t=1}^{t} \frac{n_i^t}{n_i} \left(\frac{1}{g} \sum_{i=1}^{g} w_i(t) \right)$$

Note that, according to our default normalization, the average balanced worth is set equal to one, so that we have:

$$R_i(\mathbf{A},t) = W_i\ (\mathbf{A},t) - W_i^R(\mathbf{A},t) = \left[\sum_{t=1}^{t} \frac{n_i^t}{n_i} w_i(t)\right] - 1 \qquad [6]$$

3.2 Separate evaluation by groups

A different problem regarding heterogeneous populations is that of measuring the relative performance of the different types *within* groups and providing a summary measure of their degree of diversity.

Assume, as before, that each group $i = 1,2,...,g$ consists of t different types. The outcome distribution of group *i* will be given by a matrix:

$$\mathbf{A}(i) = \left[\mathbf{a}^1(i), \mathbf{a}^2(i), ..., \mathbf{a}^t(i)\right] \quad i = 1,2,...,g$$

where $\mathbf{a}^t(i) = \left(a_{i1}^t, a_{i2}^t, ..., a_{is}^t\right)$, for *t* = 1, 2, …, t , is the vector that describes the shares of type *t* within group *i* into the different levels of achievement.

The balanced worth of each of those partitioned groups, considered in isolation, $\mathbf{w}(i) = \left(w_1(i), w_2(i), ..., w_t(i)\right)$, for $i = 1,2,...,g$, tells us about the relative performance of the types within group *i*. Depending on the problem under consideration and the nature of the types, those values may provide measures of segregation, discrimination, intergenerational progress, etc.

A real-valued measure of the degree of heterogeneity for group *i* can be obtained from the dispersion of those values. Such a measure would permit one comparing heterogeneity between groups, in terms of the dispersion of the components of the balanced worth of their constituent types. Two remarks are to be made on this respect. First, we have to take into account the differences in size of the types when defining this overall heterogeneity measure. Second, the appropriate dispersion measure may vary depending on the problem under consideration (in particular on whether we want to

attach differential weights to the relative achievements of the different types).[2]

3.3 The joint evaluation

In some cases we might be willing to perform a joint evaluation. That is, comparing all types of all groups as if they were different populations. In this case we would simply apply the balanced worth, $\mathbf{w}(\mathbf{A},t)=\left[\left(w_{11},...,w_{1t}\right),...,\left(w_{g1},...,w_{gt}\right)\right]$, to the extended problem consisting of $g \acute{} t$ sub-groups. Out of this evaluation we could recover the evaluation of the groups in terms of a weighted sum, with weights corresponding to the population shares. That is,

$$w_i(\mathbf{A},t)=\sum_{t=1}^{t}\frac{n_{it}}{n_i}w_{it} \qquad [7]$$

Note that the evaluation of group *i* in equation [7] may differ from that obtained in equation [4], even though both are weighted sums of group *i*'s types values. And it may also be different from the within group evaluation, $w_t(i)$. The reason is that $w_i(t) \neq w_{it} \neq w_t(i)$ because each evaluation provides a relative measure of goodness of type *t* of group *i with respect to different terms of comparison*. The value $w_i(t)$ is the relative evaluation of type *t* from group *i* with respect to type *t* populations of other groups. The value w_{it}, on the contrary, is the relative evaluation of type *t* from group *i* with respect to all other types no matter the groups they belong to. Finally, the value $w_t(i)$ corresponds to the relative evaluation of type *t* from group *i* with respect to all other types within this group.

We can also derive an overall evaluation of the types, given by:

$$w_t(\mathbf{A},t)=\sum_{i=1}^{g}\frac{n_{it}}{n_t}w_{it} \qquad [8]$$

This evaluation will also differ from the one obtained by averaging the $w_t(i)$ values, for the same reason explained above.

3.4 Life satisfaction in Spain revisited

Let us consider now that life satisfaction is gender dependent, thus enriching the empirical example in Section 2.2. We keep age groups as our reference groups and consider two different types within each of those groups, men and women. Table 3 provides the basic data.

Table 3 around here

Consider now the separate evaluation by types. We aim at assessing how life satisfaction varies among men by age groups, and how life satisfaction varies among women by age groups. Table 4 provides the results in terms of the balanced worth and the (normalized) mean values. Mind that, even though the table contains information about both types, we are actually presenting two independent evaluations, that for men and that for women. This implies that making comparisons by rows is meaningless, except for the coefficient of variation that shows that there is larger diversity between women than between men. We observe that life satisfaction declines with age, except for the older group of men. We also find here a much larger variability in the balanced worth than in average values for both types.

Table 4 around here

Table 5 provides the separate evaluation by groups. In spite of having a single table, we are actually presenting four separate evaluations. In this case the only meaningful comparisons are by rows (between men and women for each particular age group). Women are more satisfied with life than men for all age groups except the oldest one (this partly reflects the differences in life expectancy). We also find here that the balanced worth discriminates more than average values: the relative differences between men and women by age groups, according to the balanced worth, are 4% for the first group, 11% for the second, 2% for the third and -10% for the last one. The corresponding values for the means are 1%, 3%, 0% and -4%.

Table 5 around here

Finally, we present the results of the joint evaluation. Now each of the cells defined by age and gender is considered as a group and evaluated accordingly. Consequently, we can compare young women with old men, young women with old women, or young men with old men, say. Table 6 provides the results. Note that the inclusion of all those population subgroups changes the values of the separate evaluation by types and age groups. Yet all qualitative traits are maintained: women fare better than men in all age groups except the older one, life satisfaction declines with age except for the older men, and the balanced worth presents a much larger variability than the mean values (about four times).

Table 6 around here

4. The "balanced worth" and the "worth"

The balanced worth can be regarded as a modification of the concept of *worth*, introduced in Herrero and Villar (2013) and applied subsequently in a series of empirical problems (see below). The worth is defined as the consistent extension of the binary principle that evaluates the relative performance of two groups proportionally to their corresponding domination probabilities. That is, $v_1 / v_2 = p_{12} / p_{21}$. This extension yields the following evaluation for each group in the general case:

$$v_i = \frac{\sum_{j \neq i} p_{ij} v_j}{\sum_{j \neq i} p_{ji}}, \quad i = 1, 2, ..., g \qquad [9]$$

The obvious difference between equations [2] and [9] is that the second does not include the probability of ties in the evaluation. This makes the evaluation concentrate on the part of the distribution in which the groups differ and ignore that in which they are similar. This implies that the worth may strongly overestimate those differences when e_{ij} is large (let us remind here that $p_{ij} + p_{ji} + e_{ij} = 1$, so that equation [9] does not distribute all the probability mass between the groups whereas equation [2] does it).

The following example illustrates this feature. Suppose we are comparing two groups, *i* and *j*, whose distributions yield the following values for the corresponding domination probabilities: $p_{ij} = 0.002$, $p_{ji} = 0.001$. The worth produces the following values: $v_i = 4/3$, $v_j = 2/3$, so that distribution *i* is regarded as twice as good as distribution *j*. Yet if one computes the probability of getting agents within the same level of achievement, we find that this has probability $e_{ij} = 1 - p_{ij} - p_{ji} = 0.997$. This

strongly suggests that both distributions are practically identical and hence that the worth overestimates the relative goodness of distribution *i*. The balanced worth yields the evaluation $w_i / w_j = 1.002002$, which is much closer to what the intuition suggests.[3]

The original paper by Herrero and Villar (2013) includes three empirical applications that illustrate the working of the worth. We shall describe two of those applications and calculate the balanced worth to compare it with the worth. We choose those two applications because they involve another standard evaluation that permits a better understanding of the differences between the balanced worth and the worth.

The first application corresponds to the evaluation of human capital based on the results of the 2013 assessment of cognitive abilities of the adult population, derived from the OECD's Program for International Assessment of Adult Competence survey (PIAAC), in reading literacy.[4] The PIAAC defines six ***levels of competence***, parameterized by certain thresholds of the test scores. The exercise consists of comparing the human capital of the participating countries out of the distribution of their populations in those levels of competence.

Table 7 contains the basic data and three different evaluations: the balanced worth (BW), the worth (W) and the mean score of the PIAAC test (normalised so that the mean value is one, as in the other two cases). We find, as one should expect, that the balanced worth provides a smoother evaluation than the worth; but also that the balanced worth still discriminates much more than the average scores. The corresponding coefficients of variation are 0.23 for the balanced worth, 0.34 for the worth, and 0,04 for the average scores.

Table 7 around here

The second application deals with the evaluation of health in the former European Union (EU 15), out of the 2011 Eurostat survey on self-perceived health status. People report their perceived state of health selecting one of the five possible states: Very good, Good, Fair, Bad, and Very bad. The exercise compares the results obtained applying the worth and the conventional 5-to-1 scoring rule. The data show that health perceptions are widely different among the citizens of the European Countries, with no correlation whatsoever between self-assessed health and a standard objective measure of health, such as life expectancy at birth (a coefficient of correlation below 0.1).

Table 8 contains the distribution of health perceptions as well as the balanced worth (BW), the worth (W), and the evaluation obtained with the 1-5 scoring rule (1-5). The coefficient of variation of the balanced worth is 0.25, smaller than that of the worth (0.37) but again much more discriminating than the naïve 1-5 scoring rule (a coefficient of variation of 0.056).

Table 8 around here

5 Discussion

The balanced worth provides an index that evaluates the relative goodness of a series of outcome distributions in terms of the likelihood of getting better or equal results. The key value judgement is that of comparing pairs of groups in terms of the probability that a random extraction from one of them yields a better or equal outcome than one random extraction from the other. The balanced worth corresponds to a

consistent application of this notion for any number of groups.

There are several aspects of this evaluation method that deserve some comments in order to better understand its nature and applicability.

Categories

The balanced worth requires very little information: the matrix of relative frequencies. This is why it can be naturally applied to evaluation problems involving categorical data, as the distribution of the elements of the population into the different categories is all we need. From this it follows that the definition of those categories (how many and how inclusive they are) is key to get a sensible evaluation. Changes in the definition of those categories affect the matrix of relative frequencies and hence the final result. As all the elements within a category are indistinguishable, the more generic the category is, the less attention we pay to individual differences. And vice-versa.

In summary, the definition of the categories is a relevant modelling choice that may influence the overall evaluation exercise. In many problems those categories are clearly defined by the nature of the problem, the accepted conventions, or the data availability. In others, however, the researcher can decide rather freely on the number and definition of those categories. A sensitivity analysis with some alternative specifications is advised in that case. The easy computation of the balanced worth makes of this an immediate exercise.

Numerical variables

Nothing prevents the application of this method to address problems involving numerical variables, either discrete of continuous. The empirical illustration on life satisfaction in Spain and the evaluation of cognitive skills through the PIAAC scores

are examples of that possibility. Yet one has to be careful when dealing with numerical variables because they are to be interpreted as indexing attributes rather than as genuinely quantitative values. In particular, one has to bear in mind, that this evaluation procedure does not compute the differences in the magnitude of the achievements, but just their distribution between the ordered categories.

In those two examples (life satisfaction and PIAAC) individual answers have been grouped into a rougher set of categories. One may reasonably wonder what is the purpose of loosing information by grouping those data into broader categories when we have all the individual numerical responses. There are two main arguments to do so in cases such as those considered here. First, different parts of the outcome range may represent qualitatively different concepts. This is the case of PIAAC, where different levels of proficiency represent different competencies and the scores are used as artifacts to operationalize the definition of those levels. Second, as it is the case of subjective evaluations made with numerical scales (the example of life satisfaction), there is no guarantee that numbers mean the same for different people (your 7 and my 7 may well represent very different things). Moreover, individual scales need not be linear (i.e. an evaluation 8 need not be twice one of 4, even for a single individual). Grouping numerical answers into categories may thus helps illuminate some structural features of the groups, enhance robustness and reduce the comparability assumptions required.

Relative evaluation

The balanced worth is an index that provides a relative evaluation of the performance of a collection of groups, which means that the value attached to each group depends on all the groups with which it is compared. In the limit, the balanced worth is not defined when there is a single group involved. This property implies that

the set of groups being compared should have something in common that makes relevant analysing their relative behaviour; otherwise the evaluation will be formally correct but of no interest. Deciding the groups that enter the comparison, therefore, matters. In some problems this is rather natural (e.g. the regions of a country) whereas in others is a modelling choice. Be as it may the number and nature of the groups involved could affect the evaluation of each participant. In the language of social choice theory, this evaluation function does not satisfy the principle of "independence of irrelevant alternatives". That is, the relative evaluation of any two groups may be altered by considering or not a third party.

Scoring rules

It might be tempting to think of the balanced worth as an endogenous way of attaching weights to the different categories or levels of performance, so that the result is a sort of weighted average. This is not (and *cannot be*) the case. The worth is not a scoring rule because it cannot be identified with any method that attaches weights to the levels in the distribution being compared. In the evaluation of the research outcomes in terms of citations (evaluation problem 3 in the Introduction), one of the most controversial aspects is how to weight publications in the different categories. Take the case in which categories correspond to deciles of the impact factor. How should we weight a publication in the top decile with respect to one in the fifth one, say? The worth does not go through this path and there is no way of interpreting the evaluation as inducing a system of weights for the categories. The evaluation criterion takes a different venue that cannot be formulated in terms of weights.

Equity

One may wonder if the evaluation provided by the balanced worth involves some equity concern, in the sense that distributions with a smaller variance tend to be ranked higher. The answer is no. In particular, the distribution in which all the population is equally distributed between the different levels may be the one with the lowest score. The balanced worth makes an assessment of the relative desirability, not the relative fairness, of a set of distributions. The evaluation can be interpreted in terms of the veil of ignorance, in the following sense. An individual has to choose a society in which to live, without knowing in advance to which social group she will belong. The only information is that of the likelihood of belonging to the different categories. The choice principle is that of selecting that society in which the probability of achieving a higher category is larger. This does not mean that we cannot perform evaluation exercises from an equity perspective, provided the problem is framed conveniently.

Populations and samples

We have implicitly assumed that we our evaluation method applies to representative samples of the population, without discussing how to test this hypothesis. When computing the balanced worth, therefore, what we actually get is an estimate of the actual balanced worth. We would like to be sure that such an estimate is "good enough". There is nothing new we can offer on this respect; standard bootstrapping techniques can be used to assess the statistical significance of the results. This type of techniques seems preferable to the traditional parametric methods, unless we have an a priori idea of the class of distributions we deal with.

Examples of applications

There are several studies that have already used the worth as the key evaluation

protocol, besides those included in Herrero & Villar (2013), already mentioned. Those studies show that this evaluation method can deal with primary data of different nature, subjective or objective, quantitative or qualitative. We shall now very briefly refer to them. Since the balanced worth is a refinement of the worth, all of those applications serve to illustrate the scope of our method.

(a) Herrero, Méndez and Villar (2014) analyse the evaluation of scholastic performance using PISA data and applying inverse probability weighting (IPW) techniques to control for differences in the distribution of the determinants of the outcome variable. Computing the covariate-adjusted evaluation permits one to isolate the impact of the explanatory variables and estimate the impact of the latent variables on the relative performance.

(b) Villar (2014) deals with the study of the results of the Programme for International Assessment of Adult Competences (PIAAC), regarding Spain, in the field of *mathematical competence*. The key element consists of comparing the relative skills acquired by the different generations that compose the Spanish working age population. The study uses the distributions of the population of the different cohorts into those five levels of competence. Each cohort is divided into three sub-types according to their educational achievements (compulsory education, secondary education and university studies), in order to perform the comparative analysis.

(c) Gallen and Peraita (2015) provide an application of the worth to the analysis of corporate social responsibility (CSR) engagement in the OECD. The interest of this question derives from the observed expansion of CSR engagement of the OECD countries in recent years, a period of financial crisis.

(d) Torregrosa (2015) uses the worth to analyse the evolution of autonomic-

nationalist feelings in Spain based on opinion surveys regarding the state of Spanish Autonomous Communities carried out by Spain's Centre for Sociological Research since 1996.

(e) Albarrán *et al.* (2016) analyse the intellectual influence by countries and research fields, from a dataset consisting of 4.4 million articles published in the period 1998-2003 and indexed by Thomson Scientific, as well as the citations they received during a five-year citation window for each year in that period. Different conventional evaluation criteria are considered and confronted with the worth. Altogether, a set of ten indicators is considered and applied to a partition of the world into 39 countries and eight geographical areas.

REFERENCES

1. Arneson, R.J., 1989, Equality and equality of opportunity for welfare, **Philosophical Studies**, 56: 77-93.
2. Bellù, L.G. and Liberati, P. (2005), Social Welfare Analysis of Income Distributions Ranking Income Distributions with Crossing Generalised Lorenz Curves, Food and Agriculture Organization of the United Nations.
3. Bourguignon, F., Ferreira, F.H.G. and Leite, P.G. (2007), Beyond Oaxaca–Blinder: Accounting for differences in household income distributions, **Journal of Economic Inequality**, DOI 10.1007/s10888-007-9063-y.
4. Chakravarty, SR and J. Silber, 2007, A generalized index of employment segregation, **Mathematical Social Sciences**, 53(2): 185-195.
5. Cohen, G.A., 1989, On the currency of egalitarian justice, **Ethics**, 99(4): 906-944.
6. Crespo, J.A., Li, Y. and Ruiz–Castillo, J. (2013), The Measurement of the Effect on Citation Inequality of Differences in Citation Practices across Scientific Fields. **PLoS ONE** 8(3): e58727. doi:10.1371/journal.pone.0058727.
7. Cuhadaroglu, T. (2013), My group beats your group: evaluating non-income inequalities, w.p. School of Economics and Finances, U. Of St. Andrews.
8. Echenique, F., and Fryer, RG., 2005, On the measurement of segregation, Labor and Demography, Econ WPA.
9. Eurostat (2015), **Quality of Life. Facts and Views**, Luxemburg, Publication Office of the European Union.
10. Frankel, DM. and Volij, O. (2011), Measuring school segregation, **Journal of Economic Theory,** 146(1): 1-38.

11. Gallén, M.L. and Peraita, C. (2015), A comparison of corporate social responsibility engagement in the OECD countries with categorical data, **Applied Economics Letters**, 22 : 1005-1009.
12. Gonzalez-Diaz, J., Hendrichx, R., and Lohmann, E (2013), Paired comparison analysis: an axiomatic approach to ranking methods, **Social Choice and Welfare,** in press.
13. Grannis, R. (2002), Segregation Indices and their Functional Inputs, **Sociological Methodology**, 32 (1), 69-84.
14. Herrero, C., Méndez, I. and Villar, A. (2014), Analysis of group performance with categorical data when agents are heterogeneous: The evaluation of scholastic performance in the OECD through PISA, ***Economics of Education Review***, vol. 40 : 140-151.
15. Herrero, C. and Villar, A. (2013), On the Comparison of Group Performance with Categorical Data. **PLoS ONE** 8(12): e84784. doi:10.1371/journal.pone.0084784.
16. Herrero, C. and Villar, A. (2014), Ranking distributions of monotone attributes, EUI working paper ECO 2014/6.
17. Laband, D.N., and Piette, M.J. (1994). The relative impacts of economics journals: 1970-1990, **Journal of Economic Literature**, 32(2), 640-666.
18. Laslier, J. (1997), **Tournament solutions and majority voting,** Springer, Berlin, Heildelberg, New York.
19. Li, F., Yi, K, and Jestes, J. (2009), Ranking Distributed Probabilistic Data, ***SIGMOD'09***, June 29–July 2.
20. Lieberson, S. (1976), Rank-sum comparisons between groups. **Sociological Methodology** 7: 276–291. doi: 10.2307/270713

21. Martínez-Mekler G, Martínez RA, del Río MB, Mansilla R, Miramontes P, et al. (2009), Universality of Rank-Ordering Distributions in the Arts and Sciences. **PLoS ONE** 4(3): e4791. doi:10.1371/journal.pone.0004791.
22. OECD (2013), **OECD Skills Outlook 2013**, OECD Publishing.
23. OECD (2014), **What students know and can do. Student performance in mathematics, reading and science** (Volume I, Revised edition). OECD Publishing.
24. Palacios-Huerta, I. and Volij, O (2004), The Measurement of Intellectual Influence, **Econometrica,** 72(3): 963-977.
25. Pavot, W. & Diener, E. (2008), The satisfaction with life scale and the emerging construct of life satisfaction, **The Journal of Positive Psychology**, *3* : 137-152.
26. Pinski, G., and Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics, **Information Processing and Management**, 12(5), 297--312.
27. Reardon, S. F. and Firebaugh, G. (2002), Measures of Multi-Group Segregation, **Sociological Methodology**, 32 : 33-76.
28. Roemer, J.E., 1993, A pragmatic theory of responsibility for the egalitarian planner, **Philosophy and Public Affairs**, 22(2): 146-166.
29. Roemer, J.E., 1998, Equality of opportunity, Harvard U. Press, New York.
30. Rosvall, M. and Bergstrom, C.T. (2007), An information-theoretic framework for resolving community structure in complex networks, **Proceedings of the National Academy of Sciences**, doi/10.1073/pnas.0611034104.
31. Ruiz-Castillo, J., Albarrán, P., Herrero, C. and Villar, A. (2016), A comparison of average-based percentile rank and other citation impact indicators, mimeo.
32. Sheriff, G. and Maguire, K. (2013) Ranking Distributions of Environmental

Outcomes Across Population Groups,

33. Shorrocks, A. F. (1983), Ranking income distributions, **Economica**, 3-17.
34. Slutzki, G., and Volij, O. (2006), Scoring of web pages and tournaments, **Social Choice and Welfare,** 26 (1): 75-92.
35. Torregrosa, R. (2015), Medición y evolución del sentimiento autonómico en España, forthcoming.
36. Veenhoven, R. (1991), Is happiness relative?, **Social indicators research**, *24* : 1-34.
37. Villar, A. (2014), Education and Cognitive Skills in the Spanish Adult Population. Intergenerational Comparison of Mathematical Knowledge form PIAAC Data, **Advances in Social Sciences Research Journal**, vol. 1, nº 1, pp. 72-88.
38. Yalonetzky, G. (2012), A Dissimilarity Index of Multidimensional Inequality of Opportunity, **The Journal of Economic Inequality**, 10(3): 343-373.

ENDNOTES

(1) This is just a simplification that does not affect the reasoning and can be easily dispensed with.

(2) Let us recall here that inequality measures typically give more weight to the realizations in the lower part of the distribution. This makes sense when heterogeneity is bad but this is not always the case. For instance when comparing years of schooling across generations in a given country, one typically would like to find that the young generation has higher values than the old one, so that perfect equality is not the desideratum.

(3) Note that, for a given problem, the probability of ties will depend on the number of admissible levels defined. The difference between the balanced worth and the worth will thus be smaller the finer the grid of possible outcomes and vanishes for continuous distributions.

(4) This programme provides internationally comparable data on the cognitive skills of the population in more than twenty countries regarding reading literacy and mathematics. We have aggregated levels 5 and 5 into a single one, labelled L5*, because level 6 is extremely thin.